BMC Bioinformatics

BioMed Central

Oral presentation

Combining dissimilarity based classifiers for cancer prediction using gene expression profiles

Ángela Blanco*[1], Manuel Martín-Merino[1] and Javier De Las Rivas[2]

Address: [1]Computer Science Department, Universidad Pontificia de Salamanca, C/Compañía, 5, 37002 Salamanca, Spain and [2]Cancer Research Center (CIC-IBMCC, CSIC/USAL), Salamanca, Spain

Email: Ángela Blanco* - ablancogo@upsa.es

* Corresponding author

from Third International Society for Computational Biology (ISCB) Student Council Symposium at the Fifteenth Annual International Conference on Intelligent Systems for Molecular Biology (ISMB)
Vienna, Austria. 21 July 2007

Published: 20 November 2007

BMC Bioinformatics 2007, **8**(Suppl 8):S3 doi:10.1186/1471-2105-8-S8-S3

This abstract is available from: http://www.biomedcentral.com/1471-2105/8/S8/S3

Background

DNA Microarrays allow us to monitor the expression level of thousands of genes simultaneously across a collection of related samples. This technology has been applied to the prediction of cancer considering the gene expression profiles in both, normal and cancer samples.

Support Vector Machines (SVM) have been applied to identify cancer samples considering the gene expression levels with encouraging results. This kind of techniques are able to deal with high dimensional and noisy data which is an important requirement in our practical problem.

However, common SVM algorithms rely on the use of the Euclidean distance which does not reflect accurately the proximities among the sample profiles [1].

This feature favors the misclassification of cancer samples (false negative errors) which is a serious drawback in our application. The SVM has been extended to incorporate non-Euclidean dissimilarities [2].

Nevertheless, no dissimilarity can be considered superior to the others because each one reflects just different features of the data and misclassifies a different set of patterns.

The false negative errors of individual classifiers can be reduced by combining non-optimal classifiers [3]. To this aim, different versions of the classifier are usually built by bootstrap sampling the patterns or the features.

However, resampling techniques reduce the size of the training set increasing the bias of individual classifiers and consequently the error of the resulting combination [4].

Our approach

To avoid the bias introduced by resampling techniques, we propose a combination strategy that builds the diversity of classifiers considering a set of dissimilarities that reflect different features of the data. In order to incorporate the dissimilarities into the SVM, they are first embedded in an Euclidean space such that the inter-pattern distances reflect the original dissimilarity matrix. Next, for each dissimilarity a C-SVM is trained. Finally, the resulting classifiers are properly combined using a voting strategy. Our method is able to work directly from a dissimilarity matrix.

Experimental results

The algorithm proposed has been tested using two benchmark datasets, Leukemia [5] and Breast Cancer [6].

Table 1 shows that the combination of dissimilarities improves significantly the Euclidean distance which is usually considered by most of SVM algorithms. The algo-

Table 1:

	% Error		%False Negative	
Method	Breast	Leukemia	Breast	Leukemia
Euclidean	10.2%	6.9%	4%	6.94%
Cosine	14.2%	1.38%	4%	1.38%
Correlation	14.2%	2.7%	6.1%	2.7%
$\aleph^2$	12.2%	1.38%	4%	1.38%
Manhattan	12.2%	5.5%	4%	4.16%
Spearman	16.3%	8.3%	6.1%	5.5%
Kendall-Tau	18.3%	8.3%	6.1%	5.5%
Bagging	6.1%	2.77%	2%	1.38%
Combination	8.1%	1.38%	2%	1.38%

Experimental results for the ensemble of SVM classifiers. Classifiers based solely on a single dissimilarity and Bagging have been taken as reference.

rithm based on the combination of dissimilarities improves the best single dissimilarity which is $\aleph^2$. In breast cancer, false negative errors are significantly reduced. Experimental results are similar for the k-NN classifier.

Conclusion

In this paper, we have proposed an ensemble of classifiers based on a diversity of dissimilarities. Experimental results suggest that the method proposed improves both, misclassification errors and false negative errors of classifiers based on a single dissimilarity.

References

1. Jiang D, Tang C, Zhang A: **Cluster Analysis for Gene Expression Data: A Survey.** *IEEE Transactions on Knowledge and Data Engineering* 2004, **16(11):**1370-1386.
2. Pekalska E, Paclick P, Duin R: **A generalized kernel approach to dissimilarity-based classification.** *Journal of Machine Learning Research* 2001, **2:**175-211.
3. Kittler J, Hatef M, Duin PW, Matas J: **On Combining Classifiers.** *IEEE Transactions on Pattern Analysis and Machine Intelligence* 1998, **20(3):**226-239.
4. Valentini G, Dietterich T: **Bias-variance analysis of support vector machines for the development of svm-based ensemble methods.** *Journal of Machine Learning Research* 2004, **5:**725-775.
5. West M, *et al.*: **Predicting the Clinical Status of Human Breast Cancer by using Gene Expression Profiles.** *PNAS* 2001, **98(20):**11462-11467.
6. Golub TR, *et al.*: **Molecular Classification of Cancer: Class Discovery and Class Prediction by Gene Expression Monitoring.** *Science* 1999, **286(15):**531-537.